International Journal Information System and Computer Science (IJISCS)

Full Paper
eISSN : 2598-246X
pISSN : 2598-0793

COMPARING TOOLS PROVIDED BY PYTHON AND R FOR EXPLORATORY DATA ANALYSIS

Mahathir Rahmany[1], Abdullah Mohd Zin[2], Elankovan A. Sundararajan[3]
[1,2,3]Centre for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, 43600, Malaysia,
E-Mail: mahathir@siswa.ukm.edu.my, amzftsm@ukm.edu.my, elan@ukm.edu.my

*Corresponding author
mahathir@siswa.ukm.edu.my

Article history:
Received: 18 August 2020
Revised: 14 October 2020
Accepted: 24 October 2020

Keywords:
exploratory data analysis,
EDA,
Python,
R,
statistics.

Abstract
To uncover the insight behind the data, the comprehensive analysis is needed. Exploratory Data Analysis (EDA) is one of practical data analysis that will guide how to reveal any hidden information in the data. By doing EDA, any pattern and issue in the data will be seen and eventually will lead hypothesis. To do EDA, beside any basic statistic is needed, a good tool to simplify the analysis is also a consideration. Python and R as a famous programming language in data science world provide method to implement that analysis. This paper will show how to perform EDA by utilizing the Python and R programming.

1.0 INTRODUCTION

Data cleaning and preparation process took almost 80% from the whole process in statistical analysis [1]. In doing so, it needs a strategy to simplify this profoundly difficult process. One of well known strategy is Exploratory Data Analysis (EDA) [2], [3]. EDA is very decisive analysis in Data Science world. It is initial step [4]–[6] to thoroughly grasp the data by find out any powerful hints such as suspected pattern [7], [8], anomaly [7], and also testing hypothesis [9] and checking assumption [10] by using descriptive statistics and graphical tool [11], [12]. EDA is not one stop process but it is multiple process [13] until the data is increasingly clear what is about [7]. Knowing these any disturbance or pattern in data beforehand, ultimately would lead an actionable insight.

The main purpose of EDA is striving to derive insight from the data. To that end, EDA provide variety of means to achieve that goal. On top of this, EDA is introduced not for statistician only but for non statisticians can perform EDA too as the nature of EDA is seek and trial [7]. Even the concept of EDA has been used years ago but the research about or the research utilize the EDA concept still ever-expanding [7], [14]. Those things show how imperative the EDA.

Before EDA was introduced by Tukey [4], [13], [15], [16], Confirmatory Data Analysis (CDA) is more well known compare to EDA. However, EDA and CDA are include in practical data analysis [17]. Different from EDA which is to generate hypothesis as the result from the analysis, CDA is more to test the already generated by the EDA before [7], [18], [19]. But both of them are performed after pre-processing analysis [20]. So in this case it shows that EDA and CDA is complementary each other [21].

There are 6 six steps to perform EDA [22]. Those are (i) Go through the attributes; (ii) Analyze the univariate data; (iii) Grasp linkage among attributes; (iv) Observe suspected pattern such as missing value and non-uniform value; (v) detect outlier; and (vi) feature engineering.

i. Go through the attribute.
Understanding the attribute on the data is very important task. By knowing the attribute of the data for example which column is nominal, ordinal, discrete or continuous, the type of statistical analysis that will be performed on that column can easily to be decided [23]. In addition by going through the attribute, the data become more familiar [24].

ii. Analyze the univariate data.
Analyze for every single variable of data to understand the form of data. For example analyze the centrality or dispersion of data [22]. By analyzing each variable, any suspected pattern on the data can be spotted easily. The figure 1 below showed the classification of descriptive statistics, where it consists with central tendency and dispersion.

iii. Grasp linkage among attributes
Determine if each attribute have linkages between one another. The linkage is could be correlation, covariance, and scatterplot.

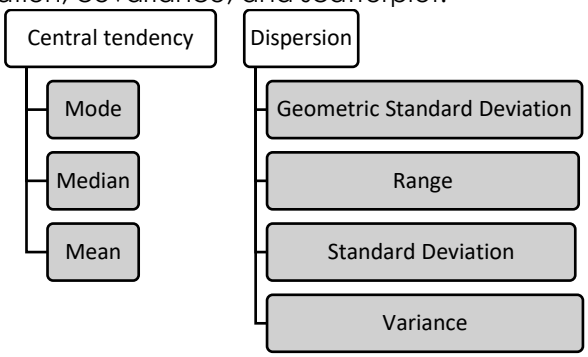

Figure 1: The example of Central Tendency and Dispersion [25], [26]

iv. Observe suspected pattern
Suspected pattern might be missed value or any pattern that look unusual or uncommon. Any action must be applied to the data in order it would not give any effect to the result analysis.

v. Detect outlier
Outlier is any data point which clearly separated from other data [27]. The outlier can influence the analysis result conclusion and as a result can generate inaccurate finding [28].

vi. Feature engineering
Feature engineering is to prepare the raw data to be applied in machine learning model by utilize the feature of the data [29].

There are many tools nowadays to perform EDA concept. They belong to the programming and nonprogramming tools. Some well-known programming tools are Python and R. Those two languages have been widely used be cutting edge programming language by researcher and scientist to do data analysis especially EDA as their prominence in the availability tool.

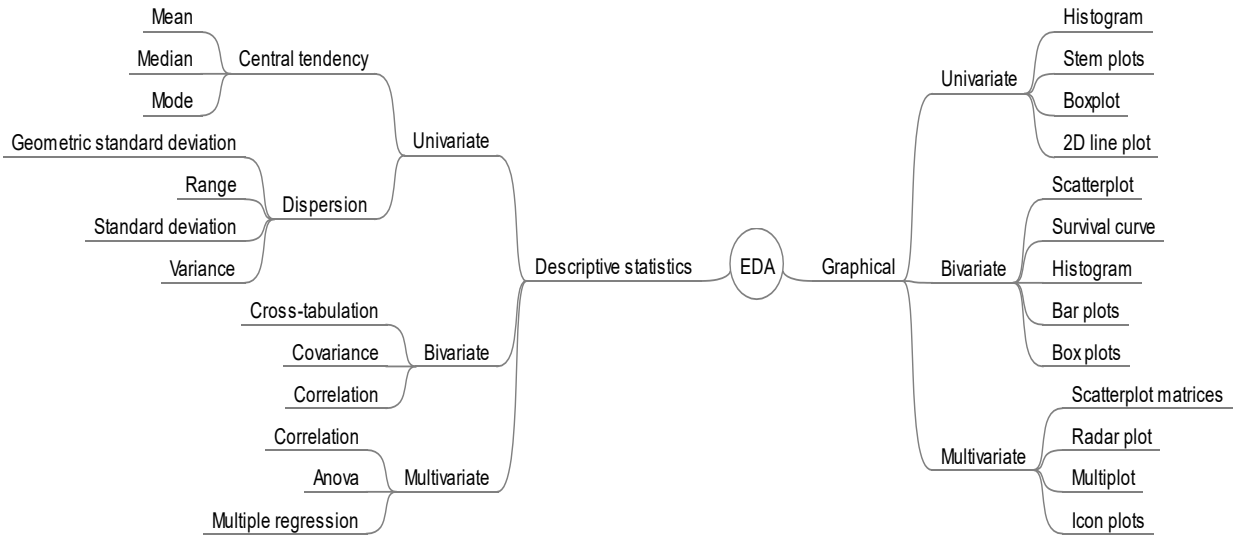

Figure 2: Technique in EDA: Descriptive and Graphical [2], [12], [16], [30]–[33]

So, in this paper will compare these two programing tools for the purpose of EDA.

2.0 TECHNIQUES IN EXPLORATORY DATA ANALYSIS

There are three types of data: (i) univariate; (ii) Bivariate; and (iii) Multivariate. Each those type of data has a different technique to applied EDA: Descriptive Statistical and graphical techniques [4]. The most common techniques is graphical [8], [34].

Figure 2 denote some type of analyzing and also type of graph or plot that commonly used for performing the EDA.

Graphical

Graphical in EDA means how to visualize the data [4] in certain ways in order to get actionable insight. From figure 2, it shows that every type of data has its own way to visualize the data even with the same way, such as like Univariate and Bivariate data share same type of visualization which is Histogram and Boxplots.

Descriptive Statistics

Descriptive statistics is the process to summarize [26] the data, be it qualitative or quantitative [35]. It is also an analysis to assess that the data is ready for the next analysis [36].

Univariate

Univariate is statistical method in analyzing the data with exclusively single or individual variable [12], [37], for example variable height only or variable weight only. As it is strictly single variable so it is very simple to analyze [34].

Bivariate

Bivariate is statistical method in analyzing two variables [30] only, for example height and weight or education level and salary. The purpose of bivariate analysis is to perceive how the independent variable can affect the dependent variable [38].

Multivariate

Multivariate is statistical method that can perform analytics on more than one relationship outright [39]. As literally multivariate data consist many variables so it become arduous to analyze [40].

Central tendency

The computation of centralized of data by evaluated the scattered of data [41], [42]. Mean, median, and mode are the example of central tendency analysis.

Dispersion

To estimate the dispersion of data [43]. The example of this kind of analysis are variance, range, and standard deviation.

Mean

The calculation of average on the data.

Median

The middle value of data when it is arranged from smallest to largest. The median is also known as 50^{th} centile or Q2 [44].

Mode
Mode is the value that most often appears in a set of data.

Standard deviation
Standard deviation (SD) is measurement the dispersion of data from the mean [31], [45].

Range
Range is one of measurement for dispersion [33] to observe the dispersion from the lowest value to highest value in a dataset.

Variance
Variance is the distance between the single data from the mean [46], [47]. So we could say that the variance is the deviation of mean squared [48].

Cross-tabulation
Cross tabulation is to perceive how two or more categorical variable interconnected [49], [50].

Covariance
Covariance is a type of descriptive statistics to see the further of the interdependence of two variables [51], [52],

Correlation
Correlation is to observe how the relationship of two variable associated [36], [51].

Histogram
Histogram is a graph to describe the frequency of continues variable [43], [51].

Stem plots
Stem plots or also known as stem and leaf plot is a graph by showing the whole data values distribution [37], [53].

Box plot
Box plot can describe either univariate or bivariate analysis. Box plot which is very good on spotting the outlier [51], is a graph to describe the variable distribution and to determinant any values that locate outside of others common value [51] or commonly known as outlier.

2D line plot
The graph that contain y and x axis where to describe array value on y axis and at x axis is equal intervals [37].

Scatterplot
The graph to visualize the relationship of two variables [54]

Bar plot
To visualize the categorical data with bar shape by directly proportional with value of the data.

3.0 TOOLS FOR EDA

Python and R are most common programming tools to perform EDA [6], [7], [55]. R and python are some of a very well known programming for data science [56]–[58].

3.1 Python

Python is one of common and famous programming language in Data Science. Even it is built not only for data analysis, but it has powerful tool (in Python world, it is known as library) to perform data analysis.

Some common libraries that is used to applied EDA in Python are Pandas [59]–[61], Numpy [60], [62] and Matplotlib [63]. Pandas is a library in Python that is used for dealing and processing the data [59], [61]. On other hand Numpy is a library to compute array based numerical application [62]. Similarly with matplotlib, the library that is focusing on making graph with many optional feature [63].

Beside those three library above there are also some libraries that really useful for EDA such as Seaborn [64], [65] and Bokeh [65].

3.2 R

R is a programming language that focus on statistics and mathematics. In term of EDA, which is belong to statistical concept, R has provided some package to deal with EDA easily.

There are some options to perform EDA in R such as using generic package or utilizing the package that is aimed for EDA only. The generic package that is commonly used in visualization EDA such as GGPLOT and Plotly. In other side there are some package that is built only to perform EDA, for example SmartEDA.

The differentiate between generic package and built-in package in exploratory data analysis in R is by using the built-in package, it provides command to perform exploratory data analysis to guide on the analysis, but not for generic one. In generic package, to perform EDA is by using common command.

3.3 Non-Programming Tools

Beside the fully programming tool as mention above, there are also some non programming tool to perform EDA such as PowerBI, Tableau, Microsoft Excel, Google Data Studio and Google Sheet.

PowerBI is a business intelligence tool that belongs to Microsoft. Beside PowerBI, Microsoft also has another tool for analyzing data which is Microsoft Excel. Beside those tools, Google also has non-programming tools to perform EDA which are known as Google Data Studio and Google Sheet. Google Data Studio and Google Sheet are an analysis tool that is run by the cloud.

4.0 COMPARISON

Comparison between the use of Python and R in carrying out EDA is done through a case study.

4.1 The case study

To show comparison to use Python and R for EDA, the example questionnaire survey data made. The data contains 16 attributes and 300 respondents. So, it means this data has 16 columns and 300 rows.

The classification of those 16 attributes are age, status and weight are numerical data; gender, status, and disease are categorical data. The Q1 until Q10 are Likert Scale data.

4.2 Using Python

a. Go through attribute.

There are two ways to go through the attribute using python: (i) describe, and (ii) head. Both will show the attribute name.

The figure 3 below show how to go through attribute using command head in python.

```
In [8]: dft.head(1)
```

Out[8]:

	Gender	age	status	weight	height
0	male	38	married	57	174

Figure 3: Attribute name

b. Analyze the univariate data

The figure 4 below show how to analyze the univariate data in python. It clearly shows that some descriptive statistic is well compute there.

```
In [11]: dft.describe()
```

Out[11]:

	age	weight
count	300.000000	300.000000
mean	52.550000	67.793333
std	10.597256	7.099669
min	35.000000	55.000000
25%	43.000000	62.000000
50%	53.000000	68.000000
75%	62.000000	73.250000
max	69.000000	79.000000

Figure 4: Analyze the univariate data

c. Grasp linkage among attributes

The figure 5 below show the visualization correlation among the attribute. In this case the library matplotlib is used to draw the figure.

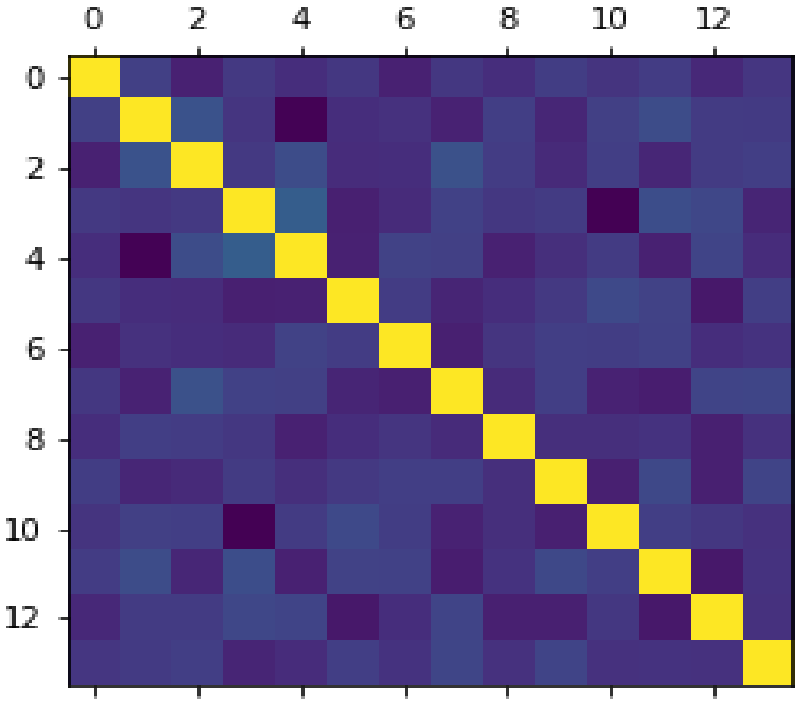

Figure 5: Correlation visualization

d. Observe suspected pattern

The figure 6 show that there is some attribute in dataset does not appropriate for next computation. For example, the Q1 attribute is supposed to be a factor, not a numeric because it is a likert scale.

```
In [4]: dft.dtypes

Out[4]: Gender      object
        age          int64
        status      object
        weight       int64
        height       int64
        disease       bool
        Q1           int64
        Q2           int64
        Q3           int64
        Q4           int64
        Q5           int64
        Q6           int64
        Q7           int64
        Q8           int64
        Q9           int64
        Q10          int64
        dtype: object
```

Figure 6: Data type

e. Detect outlier

From the figure 7, it shows that there does not have any outlier on the data.

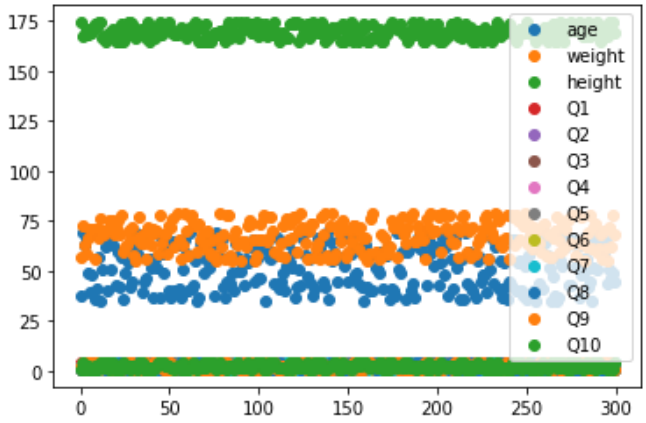

Figure 7: Detect outlier

4.3 Using R

a. Go through attribute.

There are two ways to go through the attribute: (i) head(), and (ii) summary(). Head() will show top six of the data include the name of the attribute data as shown on figure 8. In other side the summary() will show more detail about the data for example the minimal value, the maximal value, first quarter of the data, median, mean and third quarter of the.

```
> head(data)
  Gender age     status weight height disease Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10
1   male  60    married     55    166    True  3  2  2  1  3  2  1  1  4   4
2 female  47 notmarried     59    169   False  1  4  4  3  2  1  1  3  2   1
3   male  39    divorce     76    172    True  4  2  2  1  4  3  3  3  2   4
4 female  67    married     59    169   False  3  4  4  3  2  1  1  2  3   2
5   male  42 notmarried     57    166    True  3  4  1  4  1  1  2  1  1   2
6 female  63    divorce     74    167   False  4  1  2  1  3  1  3  3  1   4
```

Figure 8: head ()

From the figure 8 above we can see the dataset attribute name such as: Gender, Age, Weight, Height, Disease, Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8. Q9 and Q10.

b. Analyze the univariate data

In this example we analyze the univariate by showing the histogram of attribute age. The command to draw this graph is hist(dataset$attributename)

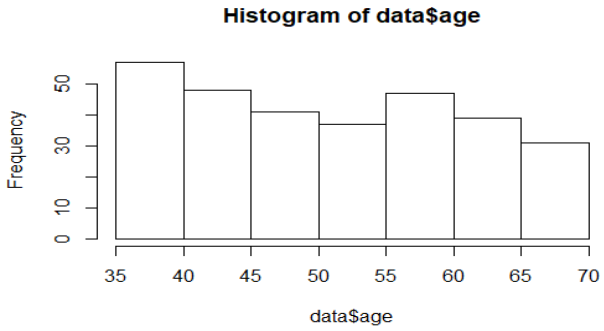

Figure 9: The histogram of attribute Age

From the figure 9 above we can see that the majority age of respondent is between 35 to 40. In the other side the age between 65 to 70 the minority age of the respondent.

c. Grasp linkage among attributes

We will see is there any relationship between the likert scale value, Q1 until Q10. From the figure 10 shows that the dispersion of the answer from the respondent is same.

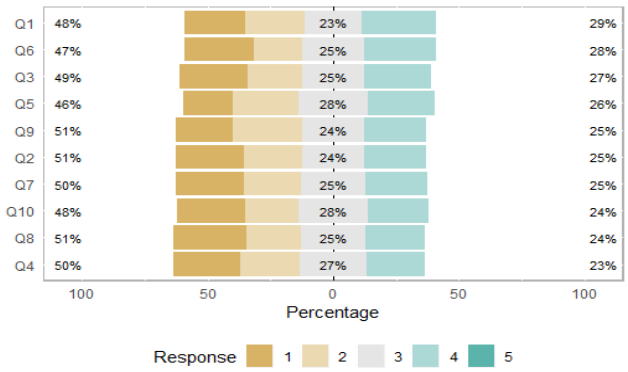

Figure 10: Bar plot for likert scale

d. Observe suspected pattern

In step four we want to check the data type of each attribute. If the data type is appropriate with the assign data.

```
$`Gender`
[1] "factor"

$age
[1] "integer"

$status
[1] "factor"

$weight
[1] "integer"

$height
[1] "integer"
```

Figure 11: Data type in data frame.

From the figure 11 above it shows that the data type for each attribute is appropriate as its assign.

e. Detect outlier

To detect outlier, we use syntax boxplot in R. From the figure 12 below, it shows there is no outlier in attribute weight.

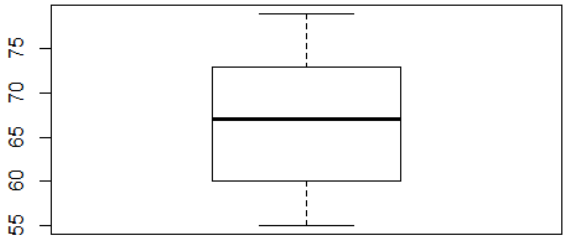

Figure 12: Detect outlier for attribute weight

5.0 RESULT AND DISCUSSION

5.1 Result

From the result of case study above, it shows that both Python and R programming language can perform the EDA smoothly. This is because both of those programming languages has provided ready-to-use syntax in their programming. For R maybe this thing is not weird because R developed to focus on statistical analysis, so it has provided any kind tool and syntax to support statistical analysis and no exception for EDA.

For python as a universal programming language, also has gear itself with statistical tool to support the EDA. Even though it does not have any special built-in package to perform EDA, as what R has, but its common package for statistical analysis is good enough to carry out EDA, for example Pandas, Numpy and Matplotlib.

5.2 Discussion

EDA is not a fixed analysis which mean that by doing the same analysis method will guide any insight in data. But EDA is iterative analysis where any possibility analysis must be performed on the data until what the data is saying is fully satisfied and understandable.

Therefore, what is shows in the Case Study section above will not fully demonstrate to find any diamond on the data. That section only want to point out that Python and R is an option that can be used to EDA beside any other options where maybe can fully accomplish the EDA for example by using Microsoft Excel, Tableau, or SPSS.

As a nature of EDA where the analysis should be done iteratively in order to find any insight information in the data, using the generic package or library is a good deal on choosing any probability that may suit to the analysis. But for the built-in package or library tied to the reserved command or analysis that has been provided. Despite that, harnessing the built-in package or library is good start to open idea on what the data talking about, then going deep into data by using the generic package or library.

6.0 CONCLUSION

EDA is an analysis to find any insight in data that will lead to hypothesis. EDA is repetitive analysis to find any information inside the data. The analysis can be done in any means without any terms that must be obeyed. However to simplify the analysis, there are some steps [22] that can be followed: (i) Go through attribute, (ii) Analyze the univariate data, (iii) Grasp linkage among attributes, (iv) Observe suspected pattern, (v) Detect outlier, and (vi) Feature engineering.

There are two type of EDA, graphical and descriptive statistics. The graphical is the way to explore data by visualizing it. In other side, to understand the data by quantitative computation is descriptive statistics. Python and R provide solution to perform EDA. With the complete syntax, exploring the data with them is very easy.

REFERENCES

[1] M. Huebner, W. Vach, dan S. [le Cessie], "A systematic approach to initial data analysis is good research practice," *J. Thorac. Cardiovasc. Surg.*, vol. 151, no. 1, hal. 25–27, 2016.

[2] M. Staniak dan P. Biecek, "The Landscape of R Packages for Automated Exploratory Data Analysis," *arXiv e-prints*, hal. arXiv:1904.02101, Mar 2019.

[3] K. A. Monsen, "Exploratory Data Analysis," in *Intervention Effectiveness Research: Quality Improvement and Program Evaluation*, Cham: Springer International Publishing, 2018, hal. 77–85.

[4] S. Putatunda, K. Rama, D. Ubrangala, dan R. Kondapalli, "SmartEDA: An R Package for Automated Exploratory Data Analysis," *arXiv Prepr. arXiv1903.04754*, 2019.

[5] C. M. Carbery, R. Woods, dan A. H. Marshall, "A New Data Analytics Framework Emphasising Pre-processing in Learning AI Models for Complex Manufacturing Systems," in *Intelligent Computing and Internet of Things*, 2018, hal. 169–179.

[6] G. L. Taboada, I. Seruca, C. Sousa, dan Á. Pereira, "Exploratory Data Analysis and Data Envelopment Analysis of Construction and Demolition Waste Management in the European Economic Area," *Sustainability*, vol. 12, no. 12, hal. 4995, 2020.

[7] A. Bezerra, I. Silva, L. A. Guedes, D. Silva, G. Leitão, dan K. Saito, "Extracting Value from

Industrial Alarms and Events: A Data-Driven Approach Based on Exploratory Data Analysis," *Sensors*, vol. 19, no. 12, hal. 2772, Jun 2019.
[8] "Understanding Clinical Data using Exploratory Analysis," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, hal. 5434–5437, Jan 2020.
[9] Z. Jones dan F. Linder, "Exploratory data analysis using random forests," in *Prepared for the 73rd annual MPSA conference*, 2015.
[10] V. Indhumathi dan N. Dharani, "Estimation of Data Analysis In R to Predict Diabetes," *Int. J. Emerg. Technol. Innov. Eng.*, vol. 6, no. 01, 2020.
[11] E. Camizuli dan E. J. Carranza, "Exploratory Data Analysis (EDA)," in *The Encyclopedia of Archaeological Sciences*, American Cancer Society, 2018, hal. 1–7.
[12] S. M. Thaung, H. M. Tun, K. K. K. Win, M. M. Than, dan A. S. S. Phyo, "Exploratory data analysis based on remote health care monitoring system by using IoT," *Communications*, vol. 8, no. 1, hal. 1–8, 2020.
[13] Y. Mao dan others, "Data Visualization in Exploratory Data Analysis: An Overview of Methods and Technologies," 2015.
[14] S. Kaleru dan S. R. Dhanikonda, "Exploratory Data Analysis and Latent Dirichlet Allocation on Yelp Database," *Int. J. Appl. Eng. Res.*, vol. 13, no. 21, hal. 15035–15039, 2018.
[15] J. W. Tukey, *Exploratory data analysis*, vol. 2. Reading, MA, 1977.
[16] K. Wongsuphasawat, Y. Liu, dan J. Heer, "Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study," *arXiv Prepr. arXiv1911.00568*, 2019.
[17] S. Thiprungsri, M. A. Vasarhelyi, A. Kogan, M. Alles, dan J. J. Ye, "Cluster analysis for anomaly detection in accounting," in *Rutgers Studies in Accounting Analytics: Audit Analytics in the Financial Industry (Rutgers Studies in Accounting Analytics)*, Emerald Publishing Limited, 2019, hal. 87–110.
[18] T. T. Allen, Z. Sui, dan K. Akbari, "Exploratory text data analysis for quality hypothesis generation," *Qual. Eng.*, vol. 30, no. 4, hal. 701–712, 2018.
[19] Z. Maznah, M. Halimah, M. Shitan, P. K. Karmokar, dan S. Najwa, "Prediction of Hexaconazole Concentration in the Top Most Layer of Oil Palm Plantation Soil Using Exploratory Data Analysis ({EDA})," *{PLOS} {ONE}*, vol. 12, no. 1, hal. e0166203, Jan 2017.
[20] X. Ma *et al.*, "Using Visual Exploratory Data Analysis to Facilitate Collaboration and Hypothesis Generation in Cross-Disciplinary Research," *ISPRS Int. J. Geo-Information*, vol. 6, no. 11, 2017.
[21] C. M. Igwenagu, "EXPLORATORY DATA ANALYSIS AND MULTIVARIATE STRATEGIES FOR REVEALING MULTIVARIATE STRUCTURES IN CLIMATE DATA," 2016.
[22] A. Ghosh, M. Nashaat, J. Miller, S. Quader, dan C. Marston, "A comprehensive review of tools for exploratory analysis of tabular industrial datasets," *Vis. Informatics*, vol. 2, no. 4, hal. 235–253, 2018.
[23] J. R. Dettori dan D. C. Norvell, "The Anatomy of Data," *Glob. Spine J.*, vol. 8, no. 3, hal. 311–313, Jan 2018.
[24] V. Cox, "Exploratory Data Analysis," in *Translating Statistics to Make Decisions : A Guide for the Non-Statistician*, Berkeley, CA: Apress, 2017, hal. 47–74.
[25] S. Deshpande, N. J. Gogtay, dan U. M. Thatte, "Measures of central tendency and dispersion," *J. Assoc. Physicians India*, vol. 64, hal. 64–66, 2016.
[26] E. G. M. Hui, "Descriptive Statistics," in *Learn R for Applied Statistics: With Data Visualizations, Regressions, and Statistics*, Berkeley, CA: Apress, 2019, hal. 87–127.
[27] C. C. Aggarwal, "An Introduction to Outlier Analysis," in *Outlier Analysis*, Cham: Springer International Publishing, 2017, hal. 1–34.
[28] M. Z. Iqbal, S. Habib, M. I. Khan, dan M. Kashif, "COMPARISON OF DIFFERENT TECHNIQUES FOR DETECTION OF OUTLIERS IN CASE OF MULTIVARIATE DATA," *Pak. J. Agri. Sci*, vol. 57, no. 3, hal. 865–869, 2020.
[29] A. Zheng dan A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, 2018.
[30] A. G. Bronevich dan J. V. de Oliveira, "On the model updating operators in univariate estimation of distribution algorithms," *Nat. Comput.*, vol. 15, no. 2, hal. 335–354, Mei 2015.
[31] R. W. Cooksey, "Descriptive Statistics for Summarising Data," in *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data*, Singapore: Springer Singapore, 2020, hal. 61–139.

[32] M. Allen, *The {SAGE} Encyclopedia of Communication Research Methods*. {SAGE} Publications, Inc, 2017.
[33] E. Mooi, M. Sarstedt, dan I. Mooi-Reci, "Descriptive Statistics," in *Springer Texts in Business and Economics*, Springer Singapore, 2017, hal. 95–152.
[34] N. Iftikhar, T. Baattrup-Andersen, F. Nordbjerg, E. Bobolea, dan P.-B. Radu, "Data Analytics for Smart Manufacturing: A Case Study," in *Proceedings of the 8th International Conference on Data Science, Technology and Applications*, 2019.
[35] S. K. Sharma, T. Kanchan, dan K. Krishan, "Descriptive Statistics," in *The Encyclopedia of Archaeological Sciences*, American Cancer Society, 2018, hal. 1–8.
[36] T. C. Guetterman, "Basics of statistics for primary care research," *Fam. Med. Community Heal.*, vol. 7, no. 2, hal. e000067, Mar 2019.
[37] M. Komorowski, D. C. Marshall, J. D. Salciccioli, dan Y. Crutain, "Exploratory Data Analysis," in *Secondary Analysis of Electronic Health Records*, Cham: Springer International Publishing, 2016, hal. 185–203.
[38] A. Bertani, G. Di Paola, E. Russo, dan F. Tuzzolino, "How to describe bivariate data," *J. Thorac. Dis.*, vol. 10, no. 2, hal. 1133–1137, Feb 2018.
[39] S. McQuitty, "The Purposes of Multivariate Data Analysis Methods: an Applied Commentary," *J. African Bus.*, vol. 19, no. 1, hal. 124–142, 2018.
[40] X. He, Y. Tao, Q. Wang, dan H. Lin, "A co-analysis framework for exploring multivariate scientific data," *Vis. Informatics*, vol. 2, no. 4, hal. 254–263, 2018.
[41] M. A. Islam dan A. Al-Shiha, "Basic Summary Statistics," in *Foundations of Biostatistics*, Singapore: Springer Singapore, 2018, hal. 39–72.
[42] U. S. Ali, "A Case Study on Teaching of Fundamental Aspects of Central Tendency by Using Classroom Activities at Secondary School Level, Karachi, Pakistan," *RADS J. Soc. Sci. Bus. Manag.*, vol. 3, no. 2, hal. 41–56, 2016.
[43] A. Gupta, P. Mishra, C. Pandey, U. Singh, C. Sahu, dan A. Keshri, "Descriptive statistics and normality tests for statistical data," *Ann. Card. Anaesth.*, vol. 22, no. 1, hal. 67, 2019.
[44] P. Shah *et al.*, "Pancreatic Glucagon secretion is severely impaired and Somatostatin secretion unchanged in patients with Hyperinsulinaemic Hypoglycaemia," in *55th Annual ESPE*, 2016, vol. 86.
[45] O. O. Mosobalaje, O. D. Orodu, dan D. Ogbe, "Descriptive statistics and probability distributions of volumetric parameters of a Nigerian heavy oil and bitumen deposit," *J. Pet. Explor. Prod. Technol.*, vol. 9, no. 1, hal. 645–661, Jun 2018.
[46] F. Kaliyadan dan V. Kulkarni, "Types of variables, descriptive statistics, and sample size," *Indian Dermatol. Online J.*, vol. 10, no. 1, hal. 82, 2019.
[47] Z. Ali dan Sb. Bhaskar, "Basic statistical tools in research and data analysis," *Indian J. Anaesth.*, vol. 60, no. 9, hal. 662, 2016.
[48] L. Igual dan S. Seguí, "Descriptive Statistics," in *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, Cham: Springer International Publishing, 2017, hal. 29–50.
[49] R. Ewing dan K. Park, *Basic Quantitative Research Methods for Urban Planners*. Taylor & Francis, 2020.
[50] S. Ismail dan S. Ahmed, "Air Pollution, its Sources and Health Effects: A Case Study of Delhi," *Res. J. Soc. Sci.*, vol. 9, no. 4, hal. 62–74, 2018.
[51] M. Sarstedt dan E. Mooi, "Descriptive Statistics," in *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, hal. 91–150.
[52] A. Indrabudiman, "Descriptive Analysis stock price with Zmijewski bankruptcy model to total assets on stock prices," *Int. J. Sci. Res. Sci. Technol. Vol.*, vol. 3, 2017.
[53] S. Boslaugh, *Encyclopedia of Epidemiology*. {SAGE} Publications, Inc., 2008.
[54] N. Fitzallen, "Interpreting Association from Graphical Displays.," *Math. Educ. Res. Gr. Australas.*, 2016.
[55] D. Borcard, F. Gillet, dan P. Legendre, "Exploratory Data Analysis," in *Numerical Ecology with R*, Cham: Springer International Publishing, 2018, hal. 11–34.
[56] M. R. M. Huddar dan R. V Kulkarni, "Role of R and Python in Data Science," *Res. JOURNEY*, hal. 32, 2018.
[57] S. K. A. Fahad dan A. E. Yahya, "Big Data Visualization: Allotting by R and Python with

GUI Tools," in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, 2018, hal. 1–8.
[58] C. D. Larose dan D. T. Larose, *Data Science Using Python and R*. Wiley, 2019.
[59] J. Bernard, "Python Data Analysis with pandas," in *Python Recipes Handbook: A Problem-Solution Approach*, Berkeley, CA: Apress, 2016, hal. 37–48.
[60] I. Meniailov, K. Bazilevych, K. Fedulov, dan S. Goranina, "Using the K-means method for diagnosing cancer stage using the Pandas library," *development*, vol. 14, hal. 15, 2019.
[61] F. Nelli, "The pandas Library---An Introduction," in *Python Data Analytics: With Pandas, NumPy, and Matplotlib*, Berkeley, CA: Apress, 2018, hal. 87–139.
[62] M. Bauer dan M. Garland, "Legate NumPy: Accelerated and Distributed Array Computing," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2019.
[63] J. Hunt, "Introduction to Matplotlib," in *Advanced Guide to Python 3 Programming*, Cham: Springer International Publishing, 2019, hal. 35–42.
[64] E. Bisong, "Matplotlib and Seaborn," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, Berkeley, CA: Apress, 2019, hal. 151–165.
[65] O. Embarak, "Data Visualization," in *Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems*, Berkeley, CA: Apress, 2018, hal. 293–342.